D192–D196 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki069

CDD: a Conserved Domain Database for protein classification

Aron Marchler-Bauer*, John B. Anderson, Praveen F. Cherukuri, Carol DeWeese-Scott, Lewis Y. Geer, Marc Gwadz, Siqian He, David I. Hurwitz, John D. Jackson, Zhaoxi Ke, Christopher J. Lanczycki, Cynthia A. Liebert, Chunlei Liu, Fu Lu, Gabriele H. Marchler, Mikhail Mullokandov, Benjamin A. Shoemaker, Vahan Simonyan, James S. Song, Paul A. Thiessen, Roxanne A. Yamashita, Jodie J. Yin, Dachuan Zhang and Stephen H. Bryant

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38 A, Room 8N805, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 22, 2004; Accepted October 5, 2004

ABSTRACT

The Conserved Domain Database (CDD) is the protein classification component of NCBI's Entrez query and retrieval system. CDD is linked to other Entrez databases such as Proteins, Taxonomy and PubMed®, and can be accessed at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=cdd. CD-Search, which is available at http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi, is a fast, interactive tool to identify conserved domains in new protein sequences. CD-Search results for protein sequences in Entrez are pre-computed to provide links between proteins and domain models, and computational annotation visible upon request. Protein–protein queries submitted to NCBI's BLAST search service at http://www.ncbi.nlm.nih.gov/BLAST are scanned for the presence of conserved domains by default. While CDD started out as essentially a mirror of publicly available domain alignment collections, such as SMART, Pfam and COG, we have continued an effort to update, and in some cases replace these models with domain hierarchies curated at the NCBI. Here, we report on the progress of the curation effort and associated improvements in the functionality of the CDD information retrieval system.

INTRODUCTION

Protein domains are distinct units of molecular evolution, usually associated with particular aspects of molecular function such as catalysis or binding. In general, they represent discrete units of three-dimensional (3D) structure. The identification of functionally characterized domains in protein sequences may give the first clues as to their molecular and cellular function.

Protein domains come in families. A dazzling array of functional diversity, and a large number of clusters grouped by obvious sequence similarity, can be reduced to anywhere between several hundred and a few thousand domain superfamilies, depending on how aggressively one groups clusters based on 3D-structural and/or functional similarities. In many cases, a single or a few search models are sufficient to uniquely identify all members of a large, diverse superfamily in a sequence database. In fact, it is possible to identify and label domains in more than two-thirds of the known protein sequences with only a few thousand domain models, as exemplified by the comprehensive collection Pfam (1). However, even a compact collection such as Pfam cannot help but create separate models for what are truly homologous families. Overlapping regions in protein sequences will sometimes be annotated by more than one model. The Conserved Domain Database (CDD) also mirrors other collections, which are largely redundant with Pfam: SMART (2) and COG (3). This, of course, aggravates the annotation problem. Users of the CD-Search resource (4) may face multiple overlapping annotations, sometimes with very similar scores but distinct functional association. This often-confusing redundancy is a necessary, but not desired property of a multiple-source collection such as CDD.

One can take certain obvious steps to reduce the redundancy, and this is what we have begun to do in CDD version v2.00. Search models are clustered based on overlapping hits in the protein database. Members of a cluster that do not significantly add to the cluster's total coverage are removed

*To whom correspondence should be addressed. Tel: +1 301 435 4919; Fax: +1 301 480 9241; Email: bauer@ncbi.nlm.nih.gov

from Entrez's default CDD collection. We have also removed search models, which annotated very few or no sequences, and search models that seem to be specific for proteins and/or domains found only in narrow phylogenetic lineages.

However, redundancy can be a good thing, if it provides more specific functional annotations, and the relationships between related models are clear and well explained to the user. There are practical limits to subdividing domain superfamilies: a large number of domain models will affect the database search time, and experimentally backed functional annotation is sparse in many cases. For CDD, we have adopted a principle of creating subfamilies only for ancient conserved domains, present in diverse organisms. We create subfamilies only when the phylogenetic distribution of member sequences suggests an origin of a domain 'orthology' group by gene duplication occurring ~0.5 Byr in the past or earlier. This principle helps us to maintain what we hope will be a uniform and understandable level of granularity. In subfamilies, we attempt to identify function from the sequence annotation and the published literature. Alignment models are kept consistent throughout superfamily hierarchies. The core model in a subfamily alignment can be mapped onto the often less extensive alignment in the 'parent' model, greatly facilitating updates to include novel representative structures and sequences.

To identify ancient subfamilies for splitting out individual search models, we perform phylogenetic analysis on the multiple sequence alignments and construct sequence trees. This procedure requires fairly accurate alignments, and frequently we do revise alignment models imported from outside sources. In alignment curation, we consider information from 3D structure and structure superposition, when possible, to define structurally conserved cores, accurately delineate domain boundaries and resolve conflicts between sequence-based alignment methods and structure superposition (5). Alignments curated at the NCBI conform to a simple block-structure, with uniformly aligned, gap-less, structurally conserved blocks separated by unaligned regions, which capture length variation.

Alignment models from both curated and imported sets are converted into position-specific scoring matrices, and the latter are assembled into search databases for use with RPS-BLAST (6).

CDD CONTENTS AND ACCESS

CDD is accessible through the Entrez data retrieval system (7), and can be queried as Entrez's 'Domains' database. Domain names and terms found in functional descriptions are indexed, and additional search capabilities are provided through reciprocal links to other Entrez resources, such as the NCBI Taxonomy Database, PubMed® and Entrez's protein database. Pre-calculated CD-Searches for proteins are recorded in the CDART database (8), which provides summaries of domain architecture for all proteins in Entrez. Pre-calculated search results are readily accessible, and provide data for protein–domain links, protein–protein links based on similar domain architecture and domain–domain links based on overlapping hit-lists.

Most of the domain models in CDD have been imported from two outside sources, Pfam and COG. CDD also contains models from SMART, and several hundred NCBI-curated domain models, identifiable by accessions starting with 'cd'. While CD-Search continues to mirror Pfam version 11.0, SMART version 4.0 and COG as individual search sets, the default 'non-redundant' CDD v2.01, as available in Entrez, currently retains only 5252 of 7255 Pfam models, 575 of 663 SMART models and 4101 of 4873 COG models. The remainder has been removed as redundant, ineffective or lineage-specific.

Search models for use with local RPS-BLAST installations, as well as CDD alignments are available at ftp://ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/. The source code for RPS-BLAST is part of the NCBI toolkit distribution, accessible at ftp://ftp.ncbi.nlm.nih.gov/toolbox.

FINDING DOMAINS IN ENTREZ

When protein queries are submitted for protein–protein BLAST® searches, they are submitted to CD-Search by default, and the resulting domain annotation is displayed graphically on the intermediate BLAST® results page. One may launch a browser window with the detailed results. Pre-calculated CD-Search results are also readily available for proteins in Entrez, following the [Domains] links.

One might, for example, study a family of plant kinesins, exemplified by gi|10130006 from *Zea mays*. CD-Search produces a graphical display as shown in the upper half of Figure 1. Two regions of the query protein receive multiple and seemingly redundant hits. The central coiled-coil region scores well with a variety of coiled-coil models contained in the uncurated subset of CDD. The C-terminal motor domain scores well with several models curated at the NCBI.

One may follow the link to the best-scoring match, 'cd01366 or KISc_C_terminal', to see the query sequence embedded in its multiple sequence alignment. What will become evident is that cd01366 is one terminal node in a larger hierarchy of related domain models, summarizing kinesin and myosin motor domains in this example. The second-best, third-best scoring hits, and so on, for the C-terminal region of gi|10130006 are to other nodes in this hierarchy. One may want to compare scores and *E*-values to understand whether the query sequence scores significantly better with one particular subgroup or not.

At the level of each individual subgroup, the similarity of the query sequence to other members of that family may be examined. We record conserved features in CD alignment models, such as active sites or binding interfaces, and their locations and residue conservation patterns may be examined in the context of the query. We provide additional annotation, such as links to literature in PubMed and links to textbooks in Entrez, so that the user can learn more about the biology of the respective families.

Building the domain family hierarchies and recording conserved features are major goals of NCBI's curation effort. We record conserved features together with evidence, such as 'structure evidence', particular 3D complexes that exemplify binding, for example, or literature citations. We also record the

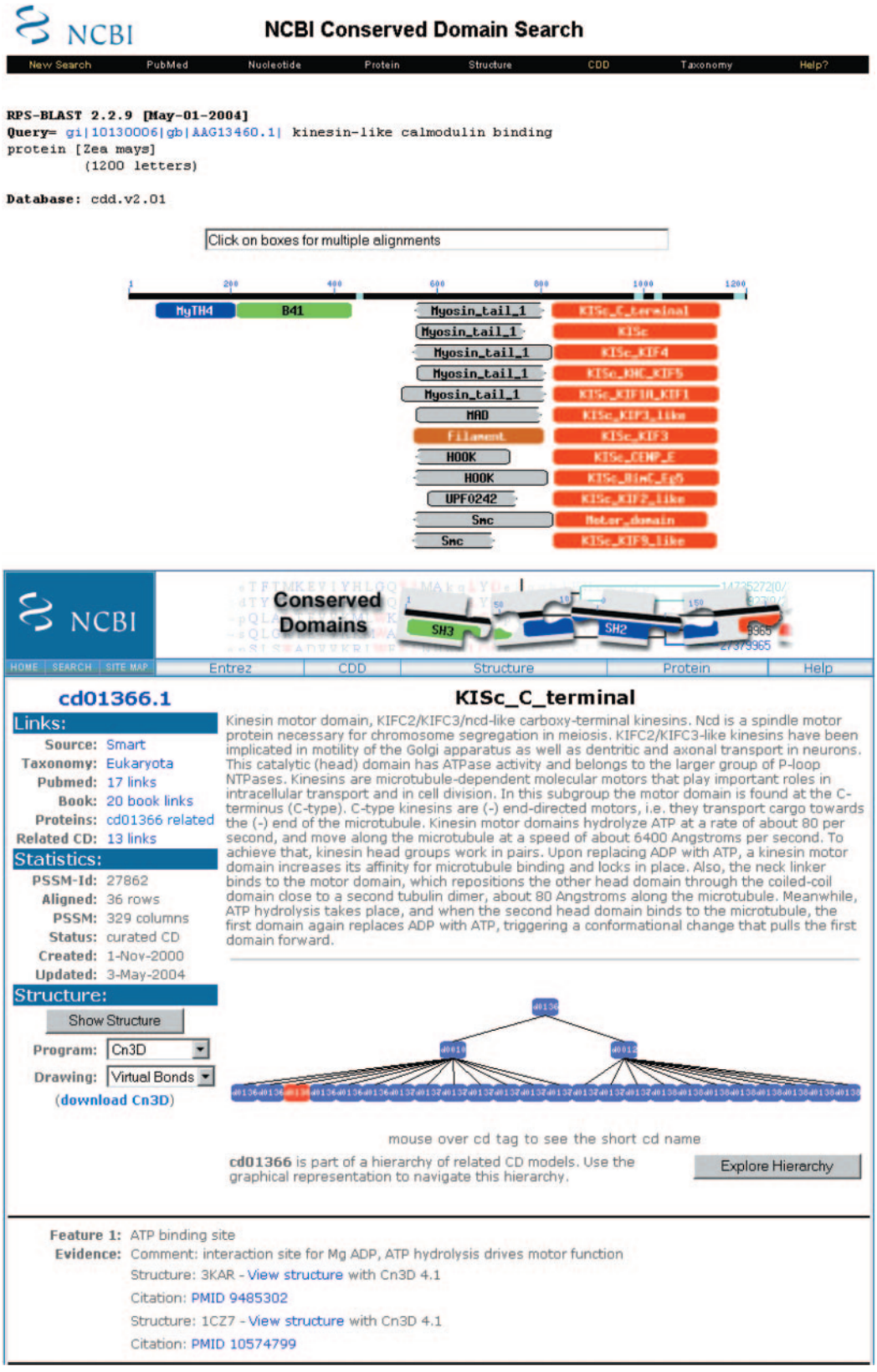

Figure 1. Pre-calculated or live CD-Search results are readily available for protein sequences in Entrez. Clicking on the colored bars will launch alignment displays that merge the query into the domain alignment model, for further analysis. Domain annotation bars with identical colors have been grouped into sets of 'related' domains, indicating that they share many of the sequence intervals hit with significant *E*-values. Annotation bars colored in gray have been classified as putative multi-domain models and are excluded from domain–domain neighboring. The lower half of the figure displays a graphical representation of a domain family hierarchy, giving the summary for one particular member (cd01366).

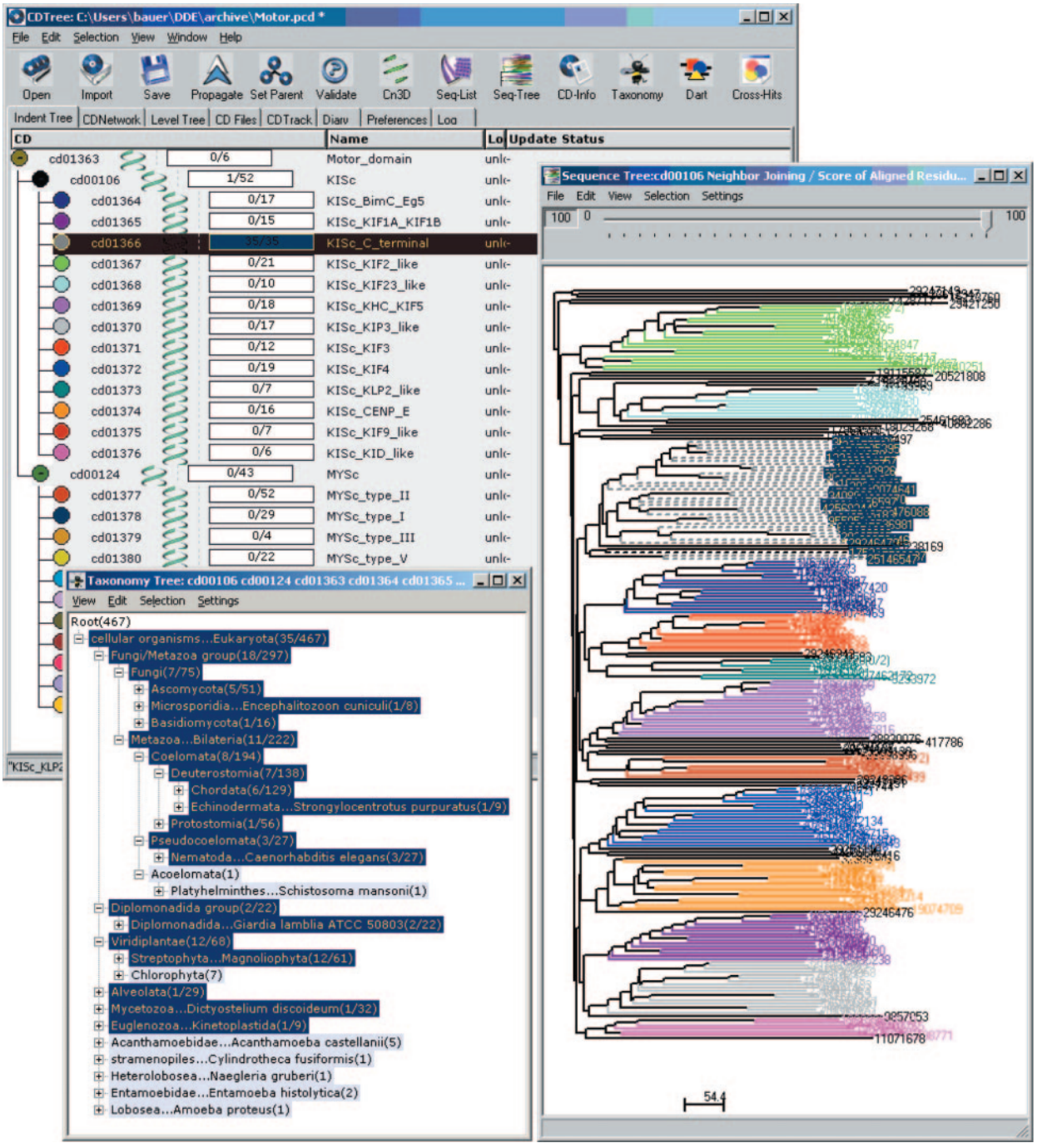

Figure 2. Subfamily hierarchy of the Myosin/Kinesin motor domains, the corresponding sequence tree and taxonomy display. One subfamily has been highlighted (KISc_C_terminal), and the highlights are reflected in both the sequence tree and taxonomy view. It is evident that members of this subfamily form a distinct node in this tree calculated by the neighbor-joining algorithm. It is also evident that members of this subfamily span a variety of taxa, suggesting that this particular type of domain was already present in their common ancestor's genome.

sequence trees used in making decisions about subfamily splits, as an evidence for the domain family hierarchy.

FUTURE DEVELOPMENTS

Beginning in 2005, we plan to distribute the software used to build and maintain these hierarchies, to serve as a helper application for the web-browser, enabling users to visualize CD family hierarchies, sequence trees and taxonomic diversity across nodes in sequence trees. Figure 2 displays a sequence-tree calculated for the Myosin/Kinesin motor domain family, which was used as an example in Figure 1.

Having access to the alignment data and analysis algorithms used by the NCBI curators should make the hierarchy editing process more transparent, should users want to investigate.

Interested users of the hierarchy editor and of Cn3D (9), the associated structure-based alignment editor, will be able to import additional sequences and examine their behavior in phylogenetic clustering.

ACKNOWLEDGEMENTS

We thank the authors of Pfam, SMART and COG, for creating invaluable resources and for helping with access to data. We also thank the NIH Intramural Research Program for support. We are grateful to the NCBI BLAST group for developing RPS-BLAST and for continuous support. Comments, suggestions and questions are welcome and should be directed to info@ncbi.nlm.nih.gov.

REFERENCES

1. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L., Studholme,D.J., Yeats,C. and Eddy,S.R. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, 138–141.
2. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res*., **32**, 142–144.
3. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N., Rao,B.S., Smirnov,S., Sverdlov,A.V., Vasudevan,S., Wolf,Y.I., Yin,J.J. and Natale,D.A. (2003) The COG database: and updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
4. Marchler-Bauer,A. and Bryant,S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res*., **32**, W327–W331.
5. Marchler-Bauer,A., Panchenko,A.R., Ariel,N. and Bryant,S.H. (2002) Comparison of sequence and structure alignments for protein domains. *Proteins*, **48**, 439–446.
6. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*., **30**, 281–283.
7. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Suzek,T.O., Tatusova,T.A. and Wagner,L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*., **32**, 35–40.
8. Geer,L.Y., Domrachev,M., Lipman,D.J. and Bryant,S.H. (2002) CDART: protein homology by domain architecture. *Genome Res*., **12**, 1619–1623.
9. Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci*., **25**, 300–302.